# CRFs-Based Chinese Patent Terminology Extraction in the Metallurgical Domain (Postprint)

**Authors:** Wang Miping, Wang Hao, Deng Sanhong, Wu Zhixiang

**Date:** 2017-10-11T00:00:00+00:00

## Abstract

【Objective】 To investigate the optimal conditions for Chinese patent term extraction models in the metallurgy domain for effectively extracting metallurgical patent terms. 【Method】 Using an imperfect core corpus and without manual annotation, a Chinese patent term recognition model for the metallurgy domain based on character role labeling was constructed using Conditional Random Fields (CRFs). The model construction process is described in detail, while focusing on comparing the influence of various CRF factors (feature combinations, character window size, etc.) on recognition performance. 【Results】Experimental results show that the combination of character sequence, level features, domain features, and temperature features, with a character window size of 3, c = 1, and f = 1, achieves a precision of 94.26%, a recall of 94.37%, and an F1 score of 94.5%. 【Limitations】 The core dictionary is imperfect, resulting in inaccurate annotation of some terms; no detailed comparison with other methods was conducted; the reliability of CRFs was not elaborated. 【Conclusion】 CRFs can effectively recognize Chinese patent terms in the metallurgy domain under appropriate combinations of roles, features, and feature templates.

## Full Text

### Extracting Chinese Metallurgy Patent Terms with Conditional Random Fields

**Wang Miping, Wang Hao, Deng Sanhong, Wu Zhixiang**
(School of Information Management, Nanjing University, Nanjing 210023, China)
(Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023, China)

## Abstract

**Objective:** This study investigates the optimal conditions for extracting Chinese patent terminology in the metallurgy domain to enable effective term recognition. **Methods:** Using an incomplete core corpus without manual annotation, we constructed a character-role labeling model for Chinese metallurgy patent term identification based on Conditional Random Fields (CRFs). We detail the model construction process and systematically compare how various CRFs factors—including feature combinations and character window size—affect recognition performance. **Results:** Experimental results demonstrate that the combination of character sequences, level features, domain features, and temperature features achieves 94.26% precision, 94.37% recall, and 94.5% F1-score when using a character window size of 3 with parameters c=1 and f=1. **Limitations:** The incomplete core dictionary leads to inaccurate labeling of some terms, and we did not conduct detailed comparisons with other methods or thoroughly discuss CRFs reliability. **Conclusions:** CRFs can effectively identify Chinese patent terms in the metallurgy domain under appropriate combinations of roles, features, and feature templates.

**Keywords:** Chinese patent terminology; Conditional Random Fields; Terminology extraction; Sequence labeling

Patents represent one of the most effective carriers of scientific and technological information, reflecting a nation' s innovation capacity. Effective patent utilization can accelerate development for both countries and enterprises. However, Chinese patent documents are unstructured texts containing numerous long terms and English abbreviations, making it difficult for researchers to directly identify core patent terminology and thereby limiting patent utilization. Term extraction is therefore crucial, as it also lays the foundation for word segmentation, syntactic analysis, and patent ontology construction.

Current Chinese terminology recognition methods fall into three categories: (1) **Rule-based methods** that rely on linguistic knowledge to design special syntactic structures or templates for matching strings. Due to language complexity, evolving grammar, and continuous emergence of new terms, these methods lack flexibility and are difficult to implement. (2) **Statistical methods** grounded in statistical theory that identify terms based on distribution patterns in corpora, commonly using frequency measures and mutual information to assess termhood. (3) **Hybrid approaches** combining rules and statistics, either applying grammatical filters after statistical processing or using rules to generate candidates before calculating statistical significance.

Conditional Random Fields (CRFs) are a typical discriminative sequence labeling model—a type of undirected graphical model that computes the joint conditional probability distribution of state labels for an entire observation sequence. Built upon Hidden Markov Models (HMM) and Maximum Entropy Models (MEM), CRFs overcome limitations such as the label bias problem in MEM by normalizing across the entire sequence. CRFs have been widely applied to Chi-

nese text processing. Deng Sanhong et al. demonstrated its effectiveness for Chinese bibliography keyword indexing. Wang Hao et al. applied it to person name recognition in web opinion analysis, showing superiority over HMM. Liu Huoyu et al. used it for paragraph segmentation, finding better performance than MEM at the cost of higher time complexity. However, CRFs applications to patent terminology remain limited. Li Peng et al. proposed a rule-integrated method based on CRFs for patent summary extraction but achieved below 50% precision, recall, and F1-score with labor-intensive rule writing. Liu Hui et al. achieved 80% precision for communication domain terms using character-based sequence labeling with CRFs, but their manual annotation was time-consuming and they did not discuss feature selection for broader applications. Huang Shaoshan et al. extracted technical and efficacy information from English patent abstracts with approximately 40% precision. Li Hongzheng et al. recognized Chinese patent prepositional phrases with over 90% accuracy, but focused on linguistic part-of-speech features with limited practical application.

This study uses Chinese patent titles from the metallurgy domain as corpus material. Through automatic character role and feature annotation using a core corpus, we establish an automatic Chinese metallurgy term extraction model using CRFs, exploring optimal recognition conditions by adjusting experimental parameters.

## 2. CRFs-Based Character-Role Patent Term Recognition Model

The model comprises three components: character role definition, feature and role annotation, and feature template construction. Role and feature annotation constitute text labeling, where role annotation relies on a core vocabulary library for term mapping and restoration, while feature selection depends on specific corpus characteristics to assist term identification. Feature templates control factors such as feature quantity and character window size. Together, these elements form the CRFs input.

### 2.1 Character-Role Labeling Model

As shown in [Figure 1: see original paper], the model consists of corpus generation and sequence labeling. In corpus generation: we first construct a Chinese core vocabulary library for the iron and steel metallurgy domain containing 6,467 domain terms, chemical elements, and common patent words sourced from websites, professional dictionaries, and domain experts. Patent titles are then segmented into character sequences (including Chinese characters and continuous letter/digit strings) and annotated with character roles. The character and role sequences combine to form learning corpora. In sequence labeling: external features such as transliteration and surname indicators extend the observation sequences. These features combine with feature templates through the CRFs algorithm to generate sequence labeling models. We test various observation

sequence values, feature sets, and role sets to identify optimal modeling conditions. Test corpora with only observation sequences generate role sequences through the trained model, from which domain terms are extracted based on predefined character roles.

Training and test corpora were downloaded from the China National Intellectual Property Administration patent search platform, comprising 7,597 Chinese patents related to the metallurgy domain. Titles served as experimental texts, with the first 1,000 used for testing and the remainder for training. CRF++0.58 was employed for experiments.

### 2.2 Definition and Annotation of Patent Term Roles and Features

Character roles are annotation markers for observations, while features extend contextual characteristics. Character sequences and extended features jointly determine the role a character exhibits.

**(1) Character-Role Space Model Definition**
Roles function in two aspects: during corpus generation, character sequences are annotated based on the core vocabulary library; during sequence labeling, they affect model generation and directly impact extraction accuracy during term restoration. We defined eight roles, as shown in .

**(2) Feature Definition**
Features extend contextual information to improve testing accuracy and depend on specific corpus characteristics. Analysis of metallurgy domain texts revealed: frequent chemical elements (e.g., aluminum, iron, manganese) including character strings like Fe, Q235, NbCFe-Mn-Si; numerous category words (e.g., process, device, equipment, system); and high-frequency temperature-related terms (e.g., fire, hot, cold). Feature definitions are summarized in .

In , vertical dashed lines represent longitudinal sequence combination constraints, including long-distance context and local context. The latter uses a character window centered on the current character, commonly 3- or 5-character windows. The example shows a 5-character window, though subsequent experiments compare both sizes. Horizontal dashed lines represent transverse sequence combination constraints. Continuous Arabic numerals and English letters are treated as single characters.

**(3) Generation of Character, Role, and Extension Sequences**
The annotation algorithm first splits sentences into individual characters, merging continuous letters or digits into single units stored in the first column as character sequences. It then checks for core vocabulary, marking roles and mapping them to character sequences. Finally, it annotates digit/letter strings and non-terms. Extension sequences are annotated based on whether individual characters appear in corresponding corpora.

### 2.3 Feature Template Construction

Feature templates describe features used during training and testing. Each line represents a template using the macro %x[Row, Col], where Row indicates relative position and Col indicates absolute column position. In , n represents Row (0=current, -1=previous, 1=next), while n-gram indicates multi-feature relationships. We designed ten templates to explore different feature combinations, as shown in .

Templates TMPT0 through TMPT5 and TMPT8 progressively expand features. TMPT5 and TMPT6 compare 3- versus 5-character windows. TMPT4 and TMPT7 examine the impact of previous character role constraints.

## 3. Experimental Analysis of Patent Term Character-Role Labeling Model

During extraction, terms are restored through character-role space mapping. As defined in , B marks term beginnings, E marks endings, and S marks single-character terms. For example, "燃气" (BE) and "淬火炉" (BME) are identified as terms. "种" serves as the character preceding "燃气," while "大" follows "加热."Correctly identified terms include single-character terms and multi-character terms beginning with B and ending with E. Identified terms encompass all single-character terms and terms beginning with B, while annotated terms are those from the core vocabulary library.

We evaluate experiments using precision (P), recall (R), F1-score, and single-character recognition rate (SP).

### 3.1 Comparison of Different Feature Templates

Results based on templates are shown in and [Figure 2: see original paper]. Single-character recognition rates exceeded 94.5% across all templates, showing minimal variation and thus excluded from [Figure 2: see original paper].

**(1) Impact of Feature Addition**
TMPT0 through TMPT5 and TMPT8 compare feature expansion effects using a 5-character window. TMPT0, using only character sequences and roles, achieved 93.14% precision and 92.49% recall, indicating character information' s dominance. Feature expansion caused slight metric fluctuations, peaking at TMPT4 with level features added—correctly identifying 3,757 terms with 93.43% precision, 93.78% recall, and 93.61% F1-score, plus maximum single-character recognition at 95.00%. Appropriate feature extension improves recognition, while irrelevant features cause interference. Recall varied more than precision, suggesting features particularly benefit term recall.

**(2) Longitudinal, Transverse, and Previous-Role Constraints**
TMPT5 and TMPT6 examine longitudinal constraints, comparing 5- versus 3-character windows. Results showed minimal difference, though TMPT6 slightly

outperformed TMPT5, indicating that window size does not correlate linearly with performance and requires corpus-specific optimization.

TMPT5 and TMPT9 examine transverse interval feature constraints. TMPT9 removed interval constraints, achieving slightly higher metrics overall, suggesting interval constraints do not necessarily improve recognition for this corpus.

TMPT5 and TMPT7 compare previous-role constraints. TMPT7' s removal of previous-role constraints significantly reduced all metrics, demonstrating the importance of such constraints.

**(3) Feature Combination Effects**
To explore individual feature impacts, we adjusted templates using a 3-character window, removing interval constraints while retaining L-1L0, then recombined effective features.

**Single Feature Comparison**
Adding any feature improved recall, though precision sometimes decreased slightly. Level feature (G) proved most beneficial, followed by classification feature (C). Domain (K) and transliteration (Y) features showed similar moderate effects, while surname (X) and temperature (T) features performed worst.

**Feature Combination Comparison**
Combining effective features G and C with others revealed that G+C did not achieve optimal performance. Instead, G combined with other features performed better, with G+X achieving maximum recall (94.71%) and G+K achieving maximum precision (94.23%). Feature combinations do not correlate linearly with individual feature effectiveness.

### 3.2 Role Definition Expansion

Previous experiments used five roles: B, M, E, A, and S. We added two roles: P (character preceding term beginning) and Q (character following term ending). For consecutive terms like "燃气" and "加热" in , priority is given to term annotation; P or Q are used only when adjacent characters are non-terms. Using six improved templates with consistent feature combinations, results showed minimal difference initially, but performance degraded with feature addition, indicating that inappropriate role expansion hinders recognition.

### 3.3 Parameter Comparison

Based on templates with highest P, R, and F1 values (ZTKG combination), we adjusted software boundary parameter c (balancing underfitting and overfitting) and feature frequency threshold f (limiting features appearing fewer than f times). As Figure 5: see original paper shows, f=1 yielded optimal results; increasing f decreased all metrics, likely due to limited features in patent corpora causing low-frequency feature filtering to reduce correct term identification. Fig-

ure 5: see original paper shows c' s minimal overall impact, with metrics rising from 1 to 4, then declining before gradually increasing to peak at c=9.

## Conclusion

By defining various roles and features, annotating patent term corpora, and extracting terms via CRFs, we examined impacts of feature models, roles, and parameters. Results indicate: (1) Appropriate feature sequence extension aids term recognition, while irrelevant features hinder it; binary feature constraints significantly help, whereas interval constraints do not benefit this corpus. (2) Role expansion does not linearly improve performance and requires corpus-specific adjustment. (3) Parameter c has minimal overall impact, while f=1 works best for feature-sparse patent documents.

Using an incomplete core corpus for automatic annotation, training on 7,597 metallurgy titles required approximately 85.45 seconds. Under optimal role, feature, and template combinations, the model achieved over 94% precision and recall, identifying 70 correct out-of-vocabulary terms such as "预热器"(preheater), "卤化物" (halide), "电炉炼钢炉" (electric furnace steelmaking furnace), "反应剂" (reactant), "锻热" (forging heat), "均热钢" (soaking steel), and "模具炉" (mold furnace). The high accuracy and ability to identify unknown terms demonstrate superiority over HMM and rule-based methods.

Limitations include: core vocabulary substitution for manual annotation saves time but causes insufficient labeling; the corpus comprises titles rather than full texts, which are more concise and structured, contributing to high metrics. Future work should apply these findings to abstracts and full texts, inviting expert evaluation of unknown terms to maximize correct term identification with minimal time and expert cost.

## References

[1] He Yanfang. The Patent Literature Study Assist in Chinese Innovation Activities [N]. China Intellectual Property News, 2012-03-23(4).

[2] Ge Xu, Lu Baohua, Yang Xianghua, et al. Utilization of Patent Literature the Development of Science and Technology Universities [J]. Technology and Innovation Management, 2005, 26(1): 68-70.

[3] Jia Zhiqi, Shao Yuejian. Enhance Enterprises' Technological Innovative Capability Through Effective Use of Patent Documents [J]. Shanxi Science and Technology, 2008(1): 91-93.

[4] Uzunbas M G, Chen C, Metaxas D. An Efficient Conditional Random Field Approach for Automatic and Interactive Neuron Segmentation [J]. Medical Image Analysis, 2016, 27: 46-58.

[5] Zhang Leihan, Lv Xueqiang, Li Zhuo, et al. Research on Extraction Methods for Domain Ontology Terminology [J]. Journal of the China Society for Scientific

and Technical Information, 2014, 33(2): 167-174.

[6] Yuan Jinsong, Zhang Xiaoming, Li Zhoujun, et al. Survey of Automatic Terminology Extraction Methodologies [J]. Computer Science, 2015, 42(8): 7-12.

[7] Tang Qing, Lv Xueqiang, Li Zhuo, et al. Research on Domain Ontology Term Extraction [J]. New Technology of Library and Information Service, 2014(1): 43-50.

[8] Wang Hao, Liu Jianhua, Su Xinning, et al. Research on Techniques and Systems of Ontology Learning for Semantic Web [J]. New Technology of Library and Information Service, 2009(1): 64-72.

[9] Gu Jun, Wang Hao. Study on Term Extraction on the Basis of Chinese Domain Texts [J]. New Technology of Library and Information Service, 2011(4): 29-34.

[10] Hua Bolin. Extracting Information Method Term from Chinese Academic Literature [J]. New Technology of Library and Information Service, 2013(6): 68-75.

[11] Zhou H T, Chen J, Dong G M, et al. Detection and Diagnosis of Bearing Faults Using Shift-invariant Dictionary Learning and Hidden Markov Model [J]. Mechanical Systems and Signal Processing, 2016, 72-73: 65-79.

[12] Le Juan, Zhao Xi. Algorithm of Beijing Opera Organization Names Entity Recognition Based on HMM [J]. Computer Engineering, 2013, 39(6): 266-271, 286.

[13] Li Lishuang, Wang Yiwen, Huang Degen. Term Extraction Based on Information Entropy and Word Frequency Distribution Variety [J]. Journal of Chinese Information Processing, 2015, 29(1): 82-87.

[14] Lu Dawei, Song Rou. Automatic Recognition of the Absent Topics in Chinese Punctuation Clauses Based on Maximum Entropy Model [J]. Computer Engineering and Science, 2015, 37(12): 2282-2293.

[15] He Jingzhou, Wang Houfeng. Chinese Word Sense Disambiguation Based on Maximum Entropy Model with Feature Selection [J]. Journal of Software, 2010, 21(6): 1287-1295.

[16] Wang Hao, Deng Sanhong. Comparative Study on HMM and CRFs Applying in Information Extraction [J]. New Technology of Library and Information Service, 2007(12): 57-63.

[17] Song D J, Liu W, Zhou T Y et al. Efficient Robust Conditional Random Fields [J]. IEEE Transactions on Image Processing, 2015, 24(10): 3124-3136.

[18] Deng Sanhong, Wang Hao, Qin Jiahang, et al. Research on Keywords Indexing for Chinese Bibliography Based on Word Roles Annotation [J]. Journal of Library Science in China, 2012, 38(2): 38-49.

[19] Wang Hao, Su Xinning. Model for Person Name Recognition Based on Role Labeling Using CRFs and Its Application to Web Opinion Analysis [J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(1): 88-96.

[20] Liu Huoyu, Wang Dongbo, Su Xinning. Research of Paragraphs Segmentation and Elements Recognition for Academic Papers Based on Multi-features [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(4): 388-397.

[21] Li Peng, Gui Jie, Qiao Xiaodong, et al. Patent Summary Information Extraction Based on Conditional Random Fields and Rule Integrated [J]. Digital Library Forum, 2010(9): 2-6.

[22] Liu Hui, Liu Yao. Patent Term Extraction Based on Conditional Random Field [J]. Digital Library Forum, 2014(12): 46-49.

[23] Huang Shaoshan, Qiao Xiaodong, Gui Jie, et al. Research on Summary of Patent Information Extraction Based on Conditional Random Field [J]. Digital Library Forum, 2010(9): 7-12.

[24] Li Hongzheng, Jin Yaohong. Recognition of Chinese Patent Text Prepositional Phrase Based on conditional Random Field [J]. Modern Chinese, 2015(7): 120-122.

[25] Peng F, McCallum A. Infomation Extraction from Research Papers Using Conditional Random Fields [J]. Information Processing and Management, 2006, 42(4): 963-979.

**Conflict of Interest Statement:** All authors declare no conflict of interest.

**Author Contributions:** Wang Miping and Wang Hao conceived the research and designed the study. Wang Miping performed experiments, collected, cleaned, and analyzed data, and drafted the manuscript. Deng Sanhong and Wu Zhixiang revised the final version.

**Supporting Data:** Available in the journal's online version at http://www.infotech.ac.cn.
[1] Wang Miping. train.txt. Training data.
[2] Wang Miping. test.txt. Test data.

**Received:** 2016-03-01
**Revised:** 2016-03-28

*Note: Figure translations are in progress. See original paper for figures.*